事件相关电位研究的统计检验力分析：影响因素及方法

念靖晴 [1] ，陈曦 [2]， 陈芳芳 [3]，牛霞 [4]，罗禹 [1*]

[1] 贵州师范大学心理学院，贵阳，550025

[2] 上海 OPPO 公司，上海，200032

[3] 芜湖市第四人民医院，芜湖，241003

[4] 安徽医科大学，合肥，230031

摘 要

统计检验力是评估研究结果稳健性和可重复性的关键指标之一，然而在事件相关电位研究中计算和报告统计检验力的规范性和完整性仍有待加强。本文通过梳理总结事件相关电位研究中统计检验力的影响因素、方法以及应用实例等，能为研究者设计或预注册事件相关电位研究方案等阶段计算和报告统计检验力提供参考依据。

关键词 脑电 事件相关电位 统计检验力 样本量 试次数量

中图分类号：B849

[*]通讯作者：罗禹，男，博士，贵州师范大学心理学院教授，E-mail: yuluo@gznu.edu.cn

1 引言

在研究可重复性危机背景下（聂丹丹, 王浩, 罗蓉, 2016; 胡传鹏等, 2016），研究结果的稳健性（robustness）和可重复性（reproducibility）对于心理学研究的发展至关重要。统计检验力（statistical power）是评估研究结果可靠性和可重复性的关键性指标之一，决定了研究结果的置信水平（Fraley & Vazire, 2014; Schweizer & Furley, 2016）。统计检验力是指当零假设（null hypothesis）为假时，统计测验正确拒绝零假设的概率，一般用 1-β 表示，通常设置为 0.8（Jacob Cohen, 2013; J. Cohen, 1988）。在假设检验中，统计检验力分析模型的主要参数有：一般由效应量（effect size）、样本量（sample size）、I 类错误（α）和 II 类错误（β）。当确定其中 3 个参数的数值后，即可以计算出第 4 个参数的数值。以效应量 Cohen's *d* 为例，在统计检验力公式中，样本量固定且 α 固定（Cohen's 0.05）的情况下，随着统计检验力降低（即高 β 水平），效应量也将同步减小。此外，先前研究已对参数间的关系、各个参数与统计检验力的关系以及常规实验情境中的应用示例进行了充分的梳理和总结 (Sommet, Weissman, Cheutin, & Elliot, 2023; 彭凡, 张力为, 周财亮, 2023; 翟宏堃, 李强, 魏晓薇, 2022; 胡竹菁, 2010; 胡竹菁, 戴海琦, 2011, 2017; 赵礼, 王晖, 2019; Vankelecom, Loeys, & Moerkerke, 2024)，本文中将不再重复阐述。以统计检验力为视角回顾过去 60 年的研究发现，科学研究领域的统计检验力约为 24% （Smaldino & McElreath, 2016）。其中，神经科学研究领域的统计检验力在 8%~30% 范围之间（Button et al., 2013），意味着在 I 类错误为 5%的流行前提下，神经科学研究领域的 II 类错误率大约在 70%~92%之间，远远低于倡导的 II 类错误率（Cohen's 20%），可能导致大多数科学研究阴性结果是虚假的（Ioannidis, 2005; Munafò et al., 2017）。

脑电技术是认知神经科学领域中极为重要和被研究者广泛使用的研究工具之一。而在脑电研究中，因事件相关电位（Event-related potential, ERP）具有潜伏期和波形恒定的鲜明特征，一直被广泛用于研究个体的认知加工过程。然而先前元分析发现的大量 ERP 研究并未进行适宜的统计检验力分析，从而导致研究的统计检验力较低，研究的可重复性差 (Clayson, Carbine, Baldwin, & Larson, 2019)。其原因可能是与行为实验相比，脑电研究的一些特殊之处会给统计检验力分析带来额外困难。

一方面，ERP 研究一直遵循实验内部重复原则。在开展研究的过程中通常需要反复测量被试在特定条件下的反应，随后对多次测量结果进行平均。这意味着对于单个被试样本，采集到的数据实际是多试次的。然而，先前的 ERP 研究中进行统计检验分析时，研究者较多关注需要测量多少个被试（被试数量，number of subjects），在一定程度上忽略了每个被试需要完成多少个试次（试次数量，number of trials），并不加以报告（Larson & Carbine, 2017）。即使样本量取决于被试数量，不透明的试次数量也直接通过测量误差影响数据质量。具体而言，在确定试次数量时经常使用模糊的，跨研究团体变异较大的经验法则而非明确的计算公式或方法（Jensen & MacDonald, 2023; Larson & Carbine, 2017），会使得观察到的

变异可能包含更多的测量误差，使得检测到真实效应的概率（统计检验力）降低。

另一方面，相较于相对成熟的量表均值或反应时等单维数据，脑电数据作为一种特殊的多维时间序列数据，在频率、时间、电压振幅等不同数据维度之间存在着系统关系，对这些关系的探索衍生出了包括时域分析、频谱分析、时频分析等在内的多种分析技术(赵文瑞, 李陈渝, 陈军君, & 雷旭, 2020)。因此，类似于模糊的试次数量，脑电数据分析过程中研究者的实验方案（兴趣变量、实验设计、成分差异）、工具因素（通道数量，采集方案）以及预处理决策（分析技术、特征工程、变量选择）同样会引入额外的误差。这种不透明性同样会通过影响测量误差进而影响统计检验力，以致于传统的统计检验力分析方法难以准确适用。

越来越多的研究表明，在进行 ERP 研究时，综合考虑统计检验力分析的影响因素（如：被试数量、试次数量等），并进行先验分析，可以在一定程度上确保适宜的统计检验力和实验结果的稳健性，从而降低研究的可重复性危机（Clayson, Carbine, Baldwin, & Larson, 2019）。此外，随着预注册（pre-register）制度的推行，研究者在预注册报告中需要对被试数量、试次数量等影响统计检验力的研究设计要素进行明确规划，以及要对选取依据进行的充分说明(Paul, Govaart, & Schettino, 2021; 赵加伟, 夏涛, 胡传鹏, 2024)。因此，本研究通过梳理总结事件相关电位研究中统计检验力的影响因素、方法以及应用实例等，能为研究者在进行相关研究设计和/或预注册等计算和报告统计检验力时提供一定的参考依据。

2 ERP 研究中统计检验力分析的影响因素

在理想状态下进行 ERP 研究统计检验力分析时至少要考虑到实验方案、实验实施/数据质量控制（测量精度）以及数据分析等层面。其中，实验方案包括感兴趣的 ERP 成分本身的特殊之处/兴趣变量、样本量、实验设计（被试内、被试间、混合等实验设计）、研究范式、预期的 ERP 效应量大小/效应幅值等方面。实验实施/数据质量控制（测量精度）包括工具因素（通道数量、采集方案）、试次数量等方面。数据分析包括分析技术（时域分析内的相关技术）、数据预处理/特征工程、统计分析方法等方面。然而，对已有研究进行梳理后发现，实际研究中研究者主要关心以下 4 个因素对 ERP 研究统计检验力分析的影响：被试数量/样本量、试次数量、效应幅值（effect magnitude）和实验设计（study design）等（Boudewyn, Luck, Farrens, & Kappenman, 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam, Adam, Quirk, Vogel, & Awh, 2021）。

2.1 被试数量

被试数量指参与研究的受试者数量，作为统计检验力分析模型的核心参数，其数量的增加会显著提高研究统计检验力。在 ERP 研究中，小被试数量是导致低统计检验力的直接原因。在进行统计检验力分析时，相比于试次数量的增加，被试数量的增加对统计检验力水平提高的作用更大（Gibney et al., 2020）。例如：Gibney 等人（2020）研究发现，在被试间实验设计中，若每组被试数量为 10 名，则得到真实显著结果的可能性极低。

2.2 试次数量

试次数量作为统计检验力分析模型的另一个重要参数，是指研究者能够采集到符合研究需求数据时相对较少的重复测量次数。ERP 在脑电信号中相对较小，研究者一般通过平均特定事件的多个试次后将 ERP 从脑电信号中提取出来。因此，信噪比（Signal-Noise Ratio, SNR；脑电数据中信号水平与噪声水平的比值）是影响 ERP 研究统计检验力的重要因素(Clayson, Baldwin, & Larson, 2013; Kim et al., 2023; W. Zhang & Kappenman, 2024)，而信噪比会随着用于平均的试次数量的平方根的增加而提升(Boudewyn et al., 2018)。具体来说，在其它条件相同的情况下，被用于平均的试次数量越多，数据的信噪比就越高，从而提升研究的效应量和统计检验力。研究发现，在被试数量较少且效应量中等时，试次数量提高约一倍左右能有效地提升研究的统计检验力，并使其达到合适水平（Boudewyn et al., 2018)。

2.3 效应幅值

效应幅值是指以微伏为单位效应的绝对值大小。具体来说，效应幅值（μV）=|A 条件/组别下 ERP 成分的平均振幅幅值-B 条件/组别下 ERP 成分的平均振幅幅值|。研究表明，效应幅值与所需的试次数量成反比，效应幅值较大的 ERP 成分往往只需少量的试次就能得到稳定的统计检验力（Baker et al., 2021; Boudewyn et al., 2018）。例如：以被试内实验设计为例，若条件间 ERP 成分的效应幅值很大时，被试数量和试次数量的增加或减少对统计检验力的影响较小；当 ERP 成分的效应幅值在中等水平时，被试数量和试次数量的变化对统计检验力的变化有很大的影响；此外，若试次数量足够多，但 ERP 成分的效应幅值较小时，通过被试数量的增加也能够得到合适的统计检验力。

2.4 实验设计

实验设计是指实施实验处理的一个计划方案（如：被试内/被试间/混合设计等）以及与计划方案有关的统计分析（如：t 检验、方差分析、线性模型分析等）。具体来说，研究者需要在研究开始前需明确实验的处理水平。一般情况下，实验处理水平越多所需要的被试数量和试次数量就越多。如图 1 所示，在效应幅值相同的情况下，统计检验力的变化在被试内设计中取决于试次数量的变化，而在被试间设计中取决于被试数量的变化。换言之，在相同效应幅值下，被试内设计得到稳定统计检验力所需的试次数量更少。例如：在被试内设计的数据模拟研究中，试次数量加倍后可以将统计检验力至少提升 1 倍，而被试数量加倍的对统计检验力的影响则较小（Jensen & MacDonald, 2023）。

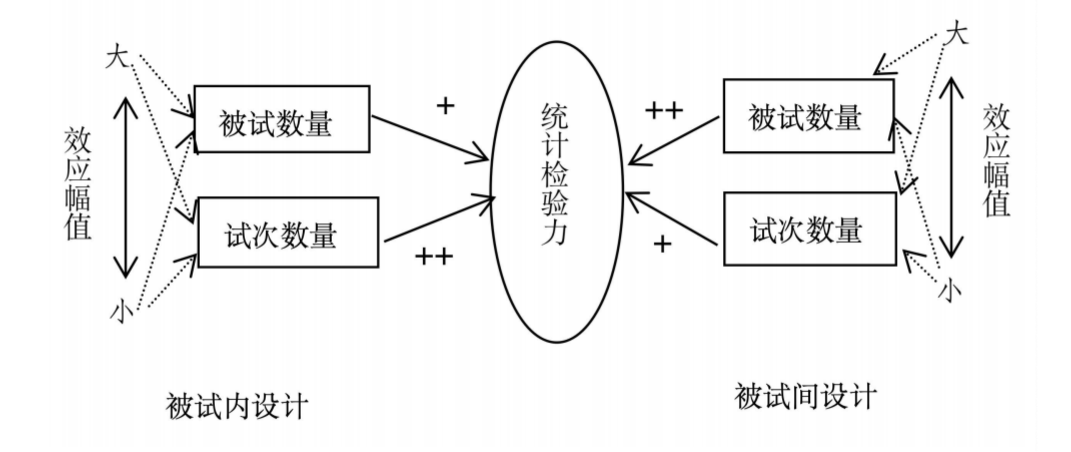


图 1 被试内实验设计和被试间实验设计模拟结果的部分关键内容示意图。被试内设计试次数量对统计检验力的影响更显著（++）。被试间设计被试数量对统计检验力的影响更显著（++）。当效应幅值存在地板效应和/或天花板效应（虚线所示）时，即效应幅值过大或过小，增加被试数量和/或试次数量对统计检验力的影响不大。图改编自 Jensen & MacDonald, 2023

3 ERP 研究中统计检验力分析方法及应用实例

统计检验力分析主要基于虚无假设显著性检验(Null Hypothesis Significance Test, NHST)，通过对核心参数进行不同组合计算，从而使统计检验力水平达到预设标准(刘玥 et al., 2024)。在实证研究中，事先科学合理地规划样本量是统计检验力分析的核心内容之一(Lakens, 2022)。因此，在 ERP 研究中进行样本量规划时，研究者需要在考虑时间和科研经费等实验成本的前提下，综合考量在不同实验设计中被试数量、试次数量、效应幅值等影响因素在统计检验力上的相互关系，从而获得样本量规划的最优解。为了获得这一最优解，研究者尝试通过事后模拟（Post-Hoc Simulations）、蒙特卡洛模拟（Monte Carlo Simulations）和检验力等高线图（Power Contours Plot）等方法分析 ERP 研究中的统计检验力。同时，上述方法各有侧重点：事后模拟主要关注在研究中获得 ERP 成分的最低试次数(Thigpen, Kappenman, & Keil, 2017)；蒙特卡洛模拟则侧重于通过灵活组合被试数量、试次数量、效应幅值和实验设计等参数以得到不同的统计检验力分析模型，并在得到模型后进行统计检验力分析(Boudewyn et al., 2018)；检验力等高线图则在充分考虑测验精度（measurement precision）和样本标准差 (Sample Standard Deviation，σ_s) 影响的前提下(Nebe et al., 2023)，动态调整被试数量和试次数量从而得到适宜的统计检验力(Baker et al., 2021)。此外，使用上述方法进行统计检验力分析时需要在相应的预实验脑电数据或者已有的脑电数据集上进行。

3.1 事后模拟

事后模拟的目的是在统计检验力、被试数量等相同的情况下，确定 ERP 研究中获得特定稳健 ERP 成分所需的最少试次数量。该方法的具体步骤是：通过进行预实验获得研究需要的 ERP 成分，然后将已经获得稳健 ERP 成分的试次数量作为总体（N），随后从总体中

抽取一定数量试次的脑电数据作为子样本（n），随后对子样本进行平均，并将平均样本数据后 ERP 成分与总体样本的 ERP 成分进行对比。不断重复上述过程，直到在子样本中得到与总体相当的 ERP 成分，并确定子样本的试次数量，该试次数量大小即为获得该 ERP 成分所需的最少试次数量。总体与子样本的相似性通过相关系数、内部一致性系数（Olvet & Hajcak, 2009; Thigpen, Kappenman, & Keil, 2017）、重测信度（Huffmeijer, Bakermans-Kranenburg, Alink, & Van IJzendoorn, 2014; Segalowitz & Barnes, 1993） 以及等值性（Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011; Pontifex et al., 2010） 等指标进行评估。例如：以内部一致性系数为例，当内部一致性系数超过 0.90 表示一致性极高，0.70-0.90 表示较高的一致性，0.50-0.70 表示中等程度的一致性，而低于 0.50 表示一致性差。Thigpen 等（2017）以内部一致性系数作为测量指标，采用事后模拟的方法对获得 P1、N1 和 P3 成分的最低试次数量进行模拟。在进行事后模拟时，抽取不同的试次数量（10~80，步长为 10）的脑电数据为子样本，随后对子样本进行叠加平均后相应成分的平均振幅值、信噪比等，并与总体（80 次左右）叠加平均后相应成分的平均振幅、信噪比进行内部一致性比较。结果发现，当子样本中的试次数量到达 40 次或以上时，子样本与总样本 ERP 成分的内部一致性系数达到 0.8 以上。结果表明，实际研究中至少 40 个试次就能得到相对稳健的 P1、N1 和 P3 成分，并不需要 80 个试次。

在应用实例方面，事后模拟被运用于 ERP 研究领域中确定错误相关负波（error-related negativity ，ERN），error positivity （Pe），N100，N200，vertex positive potential （VPP）/N170，失匹配负波（mismatch negativity ，MMN）， 反馈相关负波（feedback-related negativity ，FRN）， 晚期正成分（late positive potential ，LPP）和 P300 等 ERP 成分的试次数量（Duncan et al., 2009; Fischer, Klein, & Ullsperger, 2017; Huffmeijer et al., 2014; Jill Cohen & Polich, 1997; Larson, Baldwin, Good, & Fair, 2010; Marco-Pallares et al., 2011; Olvet & Hajcak, 2009; Pontifex et al., 2010; Rietdijk, Franken, & Thurik, 2014; Segalowitz & Barnes, 1993; Steele et al., 2016; Thigpen et al., 2017）。

事后模拟能确定获得稳健 ERP 成分所需的最少试次数量，可以在一定程度上降低相关研究的时间成本。然而，很多时候，研究的目标不仅仅在于获得稳健的 ERP 成分，可能还需要找到不同条件间的差异，但事后模拟不能量化特定的实验效应的稳定性。

3.2 蒙特卡洛模拟

相较于事后模拟，蒙特卡洛模拟能在单个模型中同时获得多个参数的统计检验力估计。在 ERP 研究中，研究者通过对被试数量、试次数量、效应幅值和实验设计等进行动态组合，从而灵活定义统计检验力分析模型。蒙特卡洛模拟的主要原理是通过指定虚拟总体（分布）以生成虚拟样本（抽样）。在关于 ERP 研究的蒙特卡洛模拟中，研究者使用预实验或者先前研究采集到的脑电数据作为指定总体，并添加了人工效应（artificial effects），从而为被试内和被试间的分析获取真实的效应幅值（Kiesel, Miller, Jolicœur, & Brisson,

2008; Smulders, 2010; Ulrich & Miller, 2001）。基本步骤为：在被试数量样本中有放回的随机抽取 n 个被试。然后在这些抽取出来的被试，他们各自的所有有效试次中随机抽取 2 组数据，每组 m 个试次。随后分别平均两组数据，之后，然后两组数据分别相加和/或相减相应的效应幅值。之后用相应的统计分析方法进行差异性检验。对于每种被试数量、试次数量和效应幅值的组合条件进行 1000 次模拟，计算每种组合条件在 1000 次的模拟中达到显著的可能性。例如：Boudewyn 等人（2018）对 ERN 成分进行蒙特卡洛模拟。其主要方法是通过让 40 名被试完成 400 个 Trials 的 Flanker 任务，并同步采集脑电数据。随后基于采集到的 40 名被试的脑电数据，采用蒙特卡洛模拟的方式模拟了 1000 个数据，最后对 1000 个数据进行分析比较。蒙特卡洛模拟结果表明， 当被试数量超过 10 个时，只需要 6 个试次就可以获得稳定的统计效力在 0.8 以上的 ERN 成分。在不同实验设计中，在不同效应幅值条件下所需要的被试数量和试次数量显著不同。在被试内实验设计中，要达到 0.8 以上的统计检验力，当被试数量均为 20 人时，在效应幅值为 4 μV 条件时，只需要 8 个试次；而当效应幅值为 2 μV 时，需要 16 个试次。在被试间实验设计中，要达到 0.8 以上的统计效力，当试次数量均为 6 个时，在效应幅值为 7 μV 条件时，只需要 16 个被试；而当效应幅值为 5 μV 条件时，需要 32 个被试。

在应用实例上，蒙特卡洛模拟分析被运用于事件相关电位研究领域中 LRP、ERN、N170、 MMN、P3、N2pc、N400、CDA、N1、Tb、P2 等 ERP 成分的统计检验力分析（Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021）。同时，为了能让研究者在实际研究中应用该方法，Hall 等人（2023）提供了在线程序 ERP Power Calculator（访问链接为：https://bradleynjack.shinyapps.io/ErpPowerCalculator/），事件相关电位研究中听觉领域的研究者可以通过选择特定的 ERP 成分（N1/Tb/P2）、试次数量（20~1000）、被试数量（10~100）、效应幅值（0~3 μV）、实验设计（被试内/被试间）、alpha 水平（0.05/0.01/0.005/0.001）等参数来计算研究的统计检验力。在视觉工作记忆领域，Ngiam 等人（2021）提供了在线程序 CDA Power Calculator（访问链接为：https://williamngiam.shinyapps.io/CDAPower/），可以通过选择感兴趣的效应（稳健 CDA 成分/记忆负荷 2 v.s 4/记忆负荷 2 v.s 6），灵活调整被试数量、干净试次数量、统计检验力等参数之间的组合来计算相应的指标。Jensen 和 MacDonald（2023）在 OSF 平台（访问链接为：https://osf.io/wv3da/）公开共享了对 LRP、ERN、N170、 MMN、P3、N2pc、N400 七个 ERP 成分通过动态组合被试数量、试次数量、效应幅值以及实验设计等参数模拟计算统计检验力的代码资源。

3.3 检验力等高线图

如前所述，除被试数量、试次数量、效应幅值和实验设计外，我们应再次关注测量精度（真实分数/总分数跨试次的均值），即通过测量误差影响统计检验力的关键指标 (Nebe et

al., 2023)。在该部分中，测量精度指重复测量具有恒定真实得分的变量并获得相似结果的能力(Cumming, 2014)，其与上述试次数量、工具、ERP 成分差异等多个因素相关。在 ERP 研究中， ERP 成分的潜伏期和波形并不具有严格的一致性和稳定性，从而导致 ERP 成分在时间、个体间和试次间的产生误差，而这一测量误差的增加会降低研究的统计检验力(Nebe et al., 2023)。Baker 等（2021）提出了检验力等高线图，在考虑个体内测量误差（within-participant variance，σ_w）和个体间测量误差（between-participants variance，σ_b）等样本标准差约束下，动态调整被试数量和试次数量并计算相应的统计检验力，直到计算的结果值达到预设标准。并将相同检验力的被试数量（N）和试次数量（k）组合成的点连成等高线，用多条等高线表示不同检验力水平(Baker et al., 2021)。在实际研究中，研究者可以通过检验力等高线在被试数量和试次数量的权衡过程中找到一个检验力的理想结合点，从而根据实际情况选取适宜的被试数量和试次数量。检验力等高线在保证样本量满足统计检验力等要求的同时又尽可能降低研究成本。例如：Baker 等人（2021）基于已有的脑电数据，对得到 P100、N600 成分被的试数量和试次数量进行重抽样，并绘制相应的统计检验力等高线。结果发现，在同等统计检验力水平下，当样本偏差较小时，P100 成分的统计检验力随被试数量和试次数量的增加而增加。N600 成分的统计检验力很大程度上取决于被试数量，而当试次数量相对较少（k＜200）时，可以通过增加试次数量来降低被试数量。

在应用实例上，统计检验力等高线图被运用于计算事件相关电位研究领域中 P100、P200、N600 等 ERP 成分以及 Alpha 频段（8~12 Hz）的被试数量和试次数量的理想结合点（Baker et al., 2021）。同时，为了方便研究者使用该方法来确定实际研究中的被试数量和试次数量，Baker 等人（2021）等人开发了在线程序 Power contour estimation（访问链接为：https://shiny.york.ac.uk/powercontours/），通过输入被试数量、试次数量、alpha 水平、均值差异、被试内标准差、被试间标准差、招募成本等参数来计算研究的统计检验力，以及实际研究中被试数量和试次数量的理想结合点。

4 ERP 研究中统计检验力分析的挑战

已有的研究系统地探讨被试数量、试次数量、效应幅值和实验设计等因素通过交互方式影响统计检验力（Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021）。但在未来的研究中还应该关注以下四点：

4.1 关注研究中可能出现的天花板效应和/或地板效应

先前研究发现，统计检验力会随着被试数量和试次数量的变化而变化，而当统计检验力出现天花板效应或地板效应时，被试数量和试次数量的变化对统计检验力的影响就微乎其微了（Boudewyn et al., 2018）。

4.2 关注事件相关电位研究中信噪比对统计检验力的影响

上述测量精度的考量只主要关注试次数量这一核心因素，但其他因素所导致的测量精

度的降低同样也不容忽视。脑电研究强调的信噪比（噪声水平），即测量精度问题，同样会导致统计检验力的降低。ERP 研究中的信噪比会受到研究范式（实验方案）、脑电数据采集（工具因素，如：不同的采集环境和设备、电阻水平等）（Kappenman & Luck, 2010; Laszlo, Ruiz-Blondet, Khalifian, Chu, & Jin, 2014; Luck & Kappenman, 2017; Picton, 2010; Puce & Hämäläinen, 2017）、特征工程/处理方法(Clayson, Baldwin, Rocha, & Larson, 2021; Delorme, 2023; G. Zhang, Garrett, & Luck, 2024a, 2024b; G. Zhang, Garrett, Simmons, Kiat, & Luck, 2023; G. Zhang & Luck, 2023; Sandre et al., 2020)、以及统计检验方法（Luck & Gaspelin, 2017）的影响。然而，蒙特卡洛模拟无法有效的模拟出每个脑电数据中真实的信噪比水平。值得强调的是，对于特征工程/处理方法，研究者主观或不经意的决策（如：采用不同的处理与分析管道等）也可能会导致假阳性结果（Luck & Gaspelin, 2017）。因此，上述列出的影响信噪比的其他因素同样是未来统计检验力研究探索的一个重要方向。

4.3 需要在更复杂的实验情境进一步验证事件相关电位研究中统计检验力的影响因素

已有的研究模拟了被试内和被试间实验设计中被试数量、试次数量以及效应幅值与统计检验力的关系（Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021）。由于脑电数据质量在研究范式、被试间、测量指标之间会有差异(G. Zhang & Luck, 2023)，因此，已有的研究结论是否适用于更复杂的实验设计（如：混合实验设计等）、分析方法（如：多因素分析、大规模单变量分析、混合效应模型等）以及不同的样本群体、实验范式中，仍需进一步验证。

4.4 在推广和应用已有研究结论时要持审慎态度

因为现有的研究结论来源是对特定 ERP 成分平均振幅幅值进行数据模拟计算后的结果。数据模拟计算的结果是一种相对理想的结果，因此可能无法推广到与模拟计算数据集有明显区别的其它数据集或数据分析方法中。如：在 ERP 研究中将成分潜伏期作为测量指标、或者在 ERP 实验中采用时频分析等方法时，除了幅度，可能还要考虑相位等对统计检验的影响。因此未来的研究中应该更加充分考虑和衡量 ERP 研究中影响统计检验力的其它潜在因素，进一步发展和推出更具有广泛适用的统计检验分析方法和计算工具。

5 ERP 研究中统计检验力分析未来发展方向与建议

在对 ERP 研究结果稳健性和可重复性受到挑战的现状的思考中，越来越多研究者开始关注低统计检验力的研究所带来的消极影响，并提出事先进行统计检验力分析来规避这一风险。在 ERP 研究中，研究的统计检验力对作者和读者都具有重要意义，如何在研究者在设计和/或预注册研究方案阶段，充分发挥 ERP 研究统计检验力分析的积极作用，不断优化研究方案，降低在低统计检验力研究上投入成本的可能性，需要各方人员共同努力。

5.1 科学合理的规划样本量

在进行实验设计时，研究者就需要以适宜的方式提前规划好样本量。对于样本量的规划方案是统计检验力分析的核心内容之一，关于样本量规划的一般原则已有前人研究总结

完善，不再赘述 (Lakens, 2022; Sommet et al., 2023)。在开展 ERP 研究时，建议使用蒙特卡洛模拟或者检验力等高线图进行样本量规划(Baker et al., 2021; Boudewyn et al., 2018; Gibney et al., 2020; Hall et al., 2023; Jensen & MacDonald, 2023; Ngiam et al., 2021)。 此外，除了通过事先规划的样本量之外，基于贝叶斯因子的序列分析也是一个重要的替代方法(郑元瑞, 胡传鹏, 2023)。

5.2 准确全面的报告内容，谨慎报告推广性结论

研究者需要认识到脑电研究，特别是 ERP 研究的重复性问题。由于认知神经科学需要考量的实验条件和工具繁复，测量模式复杂，研究者需要全面报告已知的所有实验条件和参数，为可重复性和寻找统计检验力问题提供切实的元数据(罗禹, 念靖晴, 王薇薇, 2021)。同时，研究者也需要认识到研究的局限性，谨慎报告其结论，特别是推广性结论。

5.3 采用已知的，同行认可的方案开展研究

研究者也需要认识到循证的重要性，在进行文献调研时，任何对于前人研究的改动（例如兴趣区域与通道位置）都需要提供相实的依据(Dien, 2017)，避免基于已有数据的后验分析（data-driven）。

致谢

感谢南京师范大学心理学院胡传鹏老师、加州大学戴维斯分校 Steven Luck 教授课题组张光辉博士后、贵州中医药大学吴锐老师、深圳大学心理学院张火垠同学、中国科学技术大学甘烨彤同学、温漫歆同学以及西南大学心理学部郭瑞同学在本文写作中提供的帮助和建议。

参考文献

刘玥, 徐雷, 刘红云, 韩雨婷, 游晓锋, 万志林. (2024). 置信区间宽度等高线图在线性混合效应模型样本量规划中的应用. *心理学报*, *56*(1), 124-180.

罗禹, 念靖晴, 王薇薇. (2021-03-18). 心理学脑电研究方法探讨. *中国社会科学报*, p. 8.

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题:从危机到契机. *心理科学进展*, *24*(9), 1504-1518.

胡竹菁. (2010). 平均数差异显著性检验统计检验力和效果大小的估计原理与方法. *心理学探新*, *30*(1), 68-73.

胡竹菁, 戴海琦. (2011). 方差分析的统计检验力和效果大小的常用方法比较. *心理学探新*, *31*(3), 254-259.

胡竹菁, 戴海琦. (2017). 心理学实验研究的效果大小. *心理学探新*, *37*(1), 70-77.

赵文瑞, 李陈渝, 陈军君, 雷旭. (2020). 失眠障碍与过度觉醒:来自静息态脑电和睡眠脑电的证据. *中国科学:生命科学*, *50*(3), 270-286.

赵加伟, 夏涛, 胡传鹏. (2024). 心理学研究中预注册的现状、挑战与建议. *心理科学进展*, *32*(4), 1-13.

赵礼, 王晖. (2019). 统计检验力的分析流程与多层模型示例. *心理技术与应用*, *7*(5), 276-283.

郑元瑞, 胡传鹏. (2023). 贝叶斯因子序列分析：实验设计中平衡信息与效率的新方法. *应用心理学*, 1-18.

聂丹丹, 王浩, 罗蓉. (2016). 可重复性:心理学研究不可忽视的实践. *中国临床心理学杂志*, *24*(4), 618-622.

彭凡, 张力为, 周财亮. (2023). 体育科学实验研究如何确定适宜的样本量. *上海体育学院学报*, *47*(2), 26-36.

翟宏堃, 李强, 魏晓薇. (2022). 结构方程模型统计检验力分析：原理与方法. *心理科学进展*, *30*(9), 2117-2143.

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., et al. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295-314.

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*,

55(6), e13049.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, *50*(2), 174-186. doi:10/f4h263

Clayson, P. E., Baldwin, S. A., Rocha, H. A., & Larson, M. J. (2021). The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, 245, 118712.

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437.

Cohen, J. (1988). *Statistical power analysis for tibe behavioral sciences*. Hillsdale, NJ, l, awrence Erlbanin.

Cohen, Jacob. (2013). *Statistical power analysis for the behavioral sciences* (Revised.). Academic press.

Cohen, Jill, & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology*, *25*(3), 249-255.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7-29.

Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, *13*(1), 2372.

Dien, J. (2017). Best practices for repeated measures ANOVAs of ERP data: Reference, regional channels, and robust ANOVAs. *International Journal of Psychophysiology*, *111*, 42-56.

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., et al. (2009). Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, *120*(11), 1883-1908.

Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error-and trial-number differences. *Psychophysiology*, *54*(7), 998-1009.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019.

Gibney, K. D., Kypriotakis, G., Cinciripini, P. M., Robinson, J. D., Minnix, J. A., & Versace, F. (2020). Estimating statistical power for event-related potential studies using the late positive potential. *Psychophysiology*, *57*(2), e13482.

Hall, L., Dawel, A., Greenwood, L., Monaghan, C., Berryman, K., & Jack, B. N. (2023). Estimating statistical power for ERP studies using the auditory N1, Tb, and P2 components. *Psychophysiology*, e14363.

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & Van IJzendoorn, M. H. (2014). Reliability of event-related potentials: the influence of number of trials and electrodes. *Physiology & Behavior*, *130*, 13-22.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jensen, K. M., & MacDonald, J. A. (2023). Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components. *Psychophysiology*, *60*(7), e14245.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology*, *47*(5), 888-904.

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250-274.

Kim, B., Erickson, B. A., Fernandez-Nunez, G., Rich, R., Mentzelopoulos, G., Vitale, F., et al. (2023). EEG Phase Can Be Predicted with Similar Accuracy across Cognitive States after Accounting for Power and Signal-to-Noise Ratio. *eNeuro*, *10*(9). doi:10.1523/ENEURO.0050-23.2023

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. doi:10.1525/collabra.33267

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. *Psychophysiology*, *47*(6), 1167-1171.

Larson, M. J., & Carbine, K. A. (2017). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and

recommendations for increased rigor. *International Journal of Psychophysiology*, *111*, 33-41.

Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F., & Jin, Z. (2014). A direct comparison of active and passive amplification electrodes in the same amplifier system. *Journal of Neuroscience Methods*, *235*, 298-307.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146-157.

Luck, S. J., & Kappenman, E. S. (2017). Electroencephalography and event-related brain potentials. In *Handbook of psychophysiology* (pp. 74-100). Cambridge University Press.

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852-860.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1-9.

Nebe, S., Reutter, M., Baker, D. H., Bölte, J., Domes, G., Gamer, M., et al. (2023). Enhancing precision in human neuroscience. *eLife*, *12*, e85980.

Ngiam, W. X. Q., Adam, K. C. S., Quirk, C., Vogel, E. K., & Awh, E. (2021). Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, *58*(5).

Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, 46(5), 957-961.

Paul, M., Govaart, G. H., & Schettino, A. (2021). Making ERP research more transparent: Guidelines for preregistration. *International Journal of Psychophysiology*, *164*, 52-63.

Picton, T. W. (2010). *Human auditory evoked potentials*. Plural Publishing.

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C. T., Themanson, J. R., et al. (2010). On the number of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, *47*(4), 767-773.

Puce, A., & Hämäläinen, M. S. (2017). A review of issues related to data acquisition and analysis in EEG/MEG studies. *Brain Sciences*, *7*(6), 58.

Rietdijk, W. J., Franken, I. H., & Thurik, A. R. (2014). Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. *PloS*

One, *9*(7), e102672.

Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. *International Journal of Psychophysiology*, 156, 18-39.

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, *23*, 114-122.

Segalowitz, S. J., & Barnes, K. L. (1993). The reliability of ERP components in the auditory oddball paradigm. *Psychophysiology*, *30*(5), 451-459.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Smulders, F. T. (2010). Simplifying jackknifing of ERPs and getting more out of it: retrieving estimates of participants' latencies. *Psychophysiology*, *47*(2), 387-392.

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231178728.

Steele, V. R., Anderson, N. E., Claus, E. D., Bernat, E. M., Rao, V., Assaf, M., et al. (2016). Neuroimaging measures of error-processing: Extracting reliable signals from event-related potentials and functional magnetic resonance imaging. *Neuroimage*, *132*, 247-260.

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123-138.

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*(5), 816-827.

Vankelecom, L., Loeys, T., & Moerkerke, B. (2024). How to Safely Reassess Variability and Adapt Sample Size? A Primer for the Independent Samples t Test. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231212128.

Zhang, G., Garrett, D. R., & Luck, S. J. (2024a). Optimal filters for ERP research I: A general approach for selecting filter settings. *Psychophysiology*, e14531.

Zhang, G., Garrett, D. R., & Luck, S. J. (2024b). Optimal filters for ERP research II:

Recommended settings for seven common ERP components. *Psychophysiology*, e14530.

Zhang, G., Garrett, D. R., Simmons, A. M., Kiat, J. E., & Luck, S. J. (2023-09-17). Evaluating the Effectiveness of a Common Approach to Artifact Correction and Rejection in Event-related Potential Research. bioRxiv.

Zhang, G., & Luck, S. J. (2023). Variations in ERP data quality across paradigms, participants, and scoring procedures. *Psychophysiology*, *60*(7), e14264.

Zhang, W., & Kappenman, E. S. (2024). Maximizing signal-to-noise ratio and statistical power in ERP measurement: Single sites versus multi-site average clusters. *Psychophysiology*, *61*(2), e14440.

Statistical power analysis of event-related potential studies: influencing factors and methods

Abstract

Statistical power is one of the key indicators for assessing the robustness and replicability of research results. However, the standardization and completeness of calculating and reporting statistical power in event-related potential studies still need improvement. Researchers need to pay attention to the statistical power of the study and the impact of factors such as the number of subjects, number of trials, effect magnitude, and study design on the statistical power during the design and/or pre-registration stage of the research plan,so as to continuously optimize the research plan. Reduce the possibility of investing in low-level statistical power studies.

Key words: EEG; event-related potential; statistical power; sample size; number of trials